Research Paper | Computer Science | E-ISSN No : 2454-9916 | Volume : 3 | Issue : 5 | May 2017

ANALYSIS ON TWITTER DATA OF AUTOMOBILE DOMAIN USING ONTOLOGY

Priya Gupta

M.Tech Student, Department of Computer Science & Engineering, Faculty of Engineering and Technology, MRIU at Faridabad, Haryana, India.

ABSTRACT

Corpus Data refers to as a collection of huge datasets. Sentiment analysis contributed to a popular research area for twitter. The sentiment analysis done without feature extraction fails to give the deep result about the users opinion but, features of the domain are extracted by building ontology which helps in getting the refined sentiment analysis. Ontology means a formal, explicit specification of a shared conceptualization. Conceptualization refers to an abstract model of some world phenomena. In this paper, we have used ontology to analyse the tweets to increase augmentation and efficiency of sentiments which is obtained using naïve Bayesian algorithm. The work is done in the following stages. In first stage, the tweets are extracted from Twitter4J and stored in a repository. Then sentences are extracted one by one. Sentences extracted are simplified by removing stop words and redundant words. In Second stage, the words left in the sentences are used for sense matching using WordNet-an online semantic dictionary. WordNet dictionary is used to extract features from tweets. In Third stage, Ontology is being generated by using java customized code. Crawler is being designed next to get the details about the automobile domain. The data is stored in text manner. In fourth stage, Mapping of data is done which includes mapping of ontology with the crawler data, together with ontology validation. In fifth stage, Analysis of tweets is done using ontology by applying naïve Bayesian algorithm and comparison of automobile is done which one is better and what all are the attributes that other automobile does not fall into this category.

KEYWORDS: Sentiment analysis, Ontology, Tweets, Twitter4J.

INTRODUCTION

Corpus Data is referred to as a group of huge datasets having a large quantity of information. Corpus Data is produced from varied sources like social networking sites including Facebook, Twitter etc, and therefore the data which is generated are in varied formats like structured, semi-structured or unstructured format. Twitter is considered as a common research area for sentiment analysis. For a variety of domains it offers various advantages. The sentiment analysis prepared without extraction of features fails to provide the deep result containing the user's opinion. Features of the domain can be mined by developing ontology that helps in obtaining the refined sentiment analysis. Ontology is referred as a formal, explicit specification of a shared conceptualization. Conceptualization is an abstract model of approximately world phenomena. The relationship among ontology concepts and their relations needs to be explicitly defined. Further, ontology must be machine-readable and also the ontology must capture the consensual knowledge that can be accepted by the whole community. Ontology plays a vital role for sharing and reuse of knowledge. Information organization, management as well as understanding can be improved using ontology. In the areas where dealing with a huge amount of distributed and heterogeneous computer based information, like World Wide Net, Intranet information systems, or electronic commerce, Ontology offers a significant role.

The need of using ontology is as follows. Firstly, to share the common understanding for the structure of knowledge between people and software agents. As an example, there are various different sites containing medical information or supporting medical e-commerce services. If these sites have shared and published the similar underlying ontology of the various terms they all use, then the information from different medical sites can extracted and aggregated by computer agents. The agents will make use of the aggregated knowledge to answer the user queries or as input file to different applications.

Secondly, for enabling reuse of the domain knowledge. As an example, models for several completely different domains have to represent the notion of your time. This illustration includes the notions of your time intervals, points in time, relative measures of your time, and so on. If a single group of researchers develops such ontology in detail, others will simply apply and reuse it for his/her domains. Furthermore, if we wish to build a huge ontology, one can integrate it with various existing ontologies describing parts of the massive domain.

2. LITERATURE REVIEW

Efstratios Kontopoulos and Christos Berberidis[1] have proposed the deployment of ontology-based techniques for determining the subjects discussed in tweets and breaking down each tweet into a set of aspects relevant to the subject. Their result is the assignment of a sentiment score to each distinct aspect. A baseline scenario has also been presented that deals with the domain of a popular product (smartphones) and results in comparatively evaluating the distinct features of each model series.

K. Vithiya Ruba and D. Venkatesan[2] have proposed custom sentiment analysis tool for twitter that increases the performance by increasing the overall scoring of tweets compared to the third party result, with the deployment of the ontology the subjects discussed in tweets, are analysed and the score is assigned to the feature of laptop and also the opinion words helps in identifying the mood of people and scores the tweets to the best.

Ms. Swaminarayan Priya R. et al [3] have focussed on SPARQL i.e. a W3C standard, which is a built in querying tool available with Protégé. Though it supports for RDF and RDF graphs, it still works with OWL ontology. Their paper presented the purpose of using SPARQL on the OWL ontology to verify whether it is capable enough to query OWL ontology. Their results so obtained were a proof of their support.

Ms. Swaminarayan Priya R. et al [4] have developed a sample ontology which was tested by executing semantic queries against it. All the nodes as well as each instance metadata could be searched through the SPARQL query. According to them, the main advantages of using ontology are reusability and semantically searchable. Also, in the era of Semantic Web, the ontologies have become a powerful tool for knowledge sharing and it also supports the semantic interoperability among heterogeneous distributed systems. Ontologies plus agent technologies are an essential part of the semantic web, and their combined use will make possible the sharing of heterogeneous, autonomous knowledge sources in a capable, adaptable and extensible manner. In the multi-agent system, Ontology is used to assist the interactions among different agents and improve the quality of the service provided by each agent.

R. Baracho, G. Silva, and L. Ferreira [5] has aimed to create a process of sentiment analysis based on ontologies in the automobile domain and then to develop a prototype. The process aims at making a social media analysis, identifying feelings and opinions about brands and vehicle parts. The method that guided the development process involves the construction of ontologies and a dictionary of terms that reflect the structure of the vocabulary domain. The proposed process is capable of generating information that answers questions such as: "In the opinion of the customer, which car is better: Corsa or Palio? Which one is more beautiful? Which engine is stronger?" To answer these questions by comparison, one can show a general view reflected on different social networks, indicating, for example, that for a given vehicle, a certain percentage of responses are considered positive, while for others, the percentage is considered negative.

Gopinath Ganapathy and S. Sagayaraj[6] have focused on file metadata and file content metadata can be extracted from the Application files and folders using API's. The extracted components can be stored in the Hadoop Distributed File System along with the application environment. Extracted metadata will be in XML format. XML deals with syntactic level and the Web Ontology Language (OWL) supports semantic level for the representation of domain knowledge using classes, properties and instances. This paper converts the data model elements of XML to OWL Ontology that implements the mapping the standard XML technology XSLT.

3. EXISTING APPROACH

Existing system accepts as input a tweet (or a set of tweets) regarding a specific subject and provides sentiment scores for every aspect/feature of this subject.

The methodology is divided in two phases: (a) creation of the domain ontology and (b) sentiment analysis on a set of tweets, based on the concepts and properties included in the ontology.[1]

(a) **Creation of Domain Ontology:** In order to create a domain ontology one can use Formal Concept Analysis. Formal Concept Analysis (FCA) is a mathematical data analysis theory, typically used in Knowledge Representation and Information Management. In this approach ontology visualization is done by OntoGen.

C = {Domain}; {Smartphone}

O = {set of objects}; {Nokia_lumia, Htc_one, Samsung_galaxy, apple_iphone}

A = {set of attributes}; {Display, Processor, Battery, Camera}

(b) **Sentiment Analysis on tweets:** Sentiments is analyzed using OpenDover. OpenDover is a web service that tags the opinions and sentiments detected in a textual corpus, based on the subject domain, as well as the intensity of the sentiment expression. A sentiment score s is assigned to each tweet, where s varies between [-10, 10], depending on the appreciation level of the submitted sentence. OpenDover was considered an appropriate choice for the existing approach, since it is suitable for extracting sentiment from isolated sentences. An additional advantage is OpenDover's ontology- based architecture that offers the capability of detecting each time the domain of reference, adjusting the sentiment scores accordingly.

Existing System lacks code reusability as they were working on protégé tool, for creation of ontology. In order to overcome the issues, new system with java customized code to create ontology is proposed which supports code reusability. Other fact for existing system is that it uses open dover. OpenDover has a drawback, that the exact process of extracting the sentiment from a sentence cannot be verified, the source code and methodology behind OpenDover are not publicly available. In order to overcome the issue, a custom sentiment analysis methodology in our approach is proposed.

4. PROPOSED APPROACH

The proposed approach uses an ontology which add features and meaning to the tweets and WordNet which is an online semantic dictionary, electronic lexical database of nouns, verbs, adverbs, and adjectives. WordNet helps in extraction features from tweets. In our proposed approach we are considering automobile domain. The proposed approach has the following steps:-

1. **Tweet Extraction:** Tweets are extracted from Twitter4J. Twitter4J is referred as an unofficial library of Java for the Twitter API. With Twitter4J, you can easily integrate your Java application with the Twitter service.

2. **Tweet Repository:** A repository is a central place in which an aggregation of data is kept and maintained in an organized way. The repository is made using SQL server management. The schema of tweets table which contains userid, username, tweets, after stop word tweets

3. **Stop word Removal:** Stop words are typical frequently occurring words that have little or no discriminating power, such as \a", \about", \all", etc., or other domain dependent words. Stop words are often removed. For stop words removal sentence segmentation has to be done. Then, keyword has to be extracted.

4. **Sense matching:** In this WordNet Dictionary is integrated to find the synonyms. Like for mileage, user can also milage, average etc.

5. **Ontology:** In this stage ontology is created using a java customized code. Ontology is created which involves relation between the tokens. Ontology helps in adding meaning to the tweets and relevant analysis can be obtained by removing redundant information and irrelevant information from the tweets and augmentated analysis is obtained.

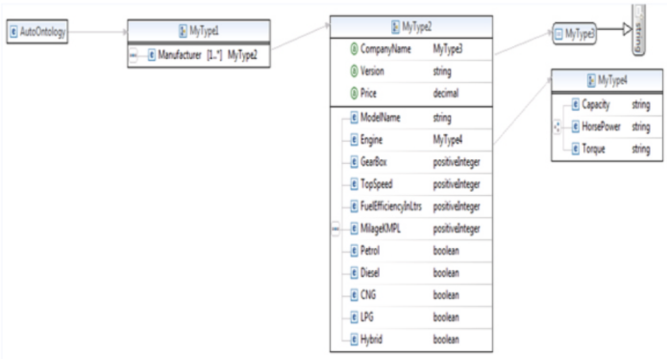

Figure 1: Ontology Design of Proposed Approach

6. **Ontology Mapping and Validation:** In this stage ontology XSD is mapped to XML. XSD contains the schema of ontology and XML contains the data instances.

7. **Comparative Analysis:** In this stage, comparative analysis is done using sentiment analysis that uses naïve Bayesian classification algorithm. We are comparing which vehicle is better over other and what are the other attributes that the other vehicle does not fall into this category on the basis of price.

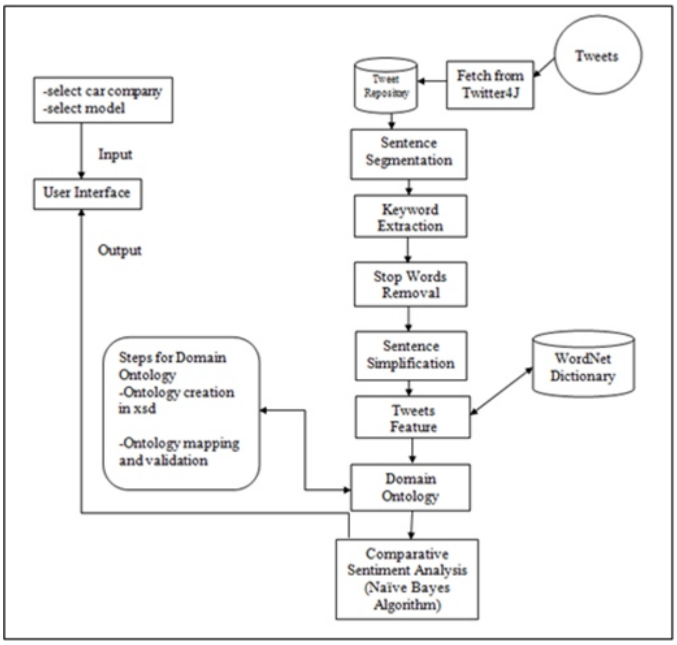

Figure 2: Architecture of proposed Approach

5. RESULTS

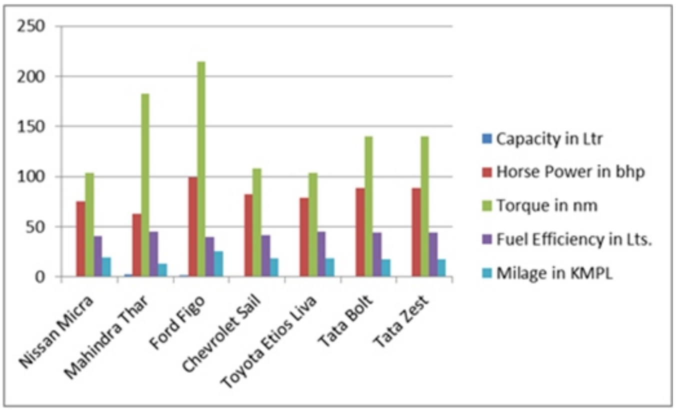

Figure 3: Car in comparision with Nissan Micra on bases of price ranging from 6.1Lakh to 6.7Lakh

The comparison of nissan micra with the other cars on the bases of price. Comparison cars include mahindra thar, ford figo, chevrolet sail, toyota etios, tata bolt and tata zest.

Ford figo is best because it is having good features like engine and mileage. We also found that Nissan micra has bad engine and poor mileage in comparison of ford figo.

Accuracy is calculated using the formula of precision i.e. the fraction of retrieved documents that are relevant to the query. For the query of Nissan micra retrieved tweets were 22 and relevant tweets were 18. Hence accuracy comes out to be81.82%. Recall is calculated using formula i.e. fraction of related tweets that are successfully retrieved. We get recall as 100%. Recall is a measure of completeness.

6. CONCLUSION AND FUTURE SCOPE

In our proposed approach, we have concluded that ontology used to analyses the tweets to increase augmentation and efficiency of sentiments obtained using is working properly. Ontology used is specification of shared conceptualization and it helps in finding hidden relationships and adds meaning to our knowledge. The main components of ontology based analysis system are tweets extraction from Twitter4J, Tweet Repository, keyword extraction, sense matching and ontology, Crawling automobile data, mapping ontology and crawler data and analysis.

In future, work can be carried forward in developing a fully automated ontology. Also, negation handling can be worked upon.

REFERENCES

[1]. Efstratios Kontopoulos and Christos Berberidis, "Expert Systems with Applications", www.elsevier.com/locate/eswa, 4065–4074, 2015.

[2]. K. Vithiya Ruba and D. Venkatesan, "Building a Custom Sentiment Analysis Tool based on an Ontology for Twitter Posts", Indian Journal of Science and Technology, Vol 8(13), July 2015.

[3]. Ms. Swaminarayan Priya R. et al, "A Comprehensive study of Query Languages for Semantic Web and retrieval of data from University Ontology Using SPARQL" , International Journal of Information and Computing Technology" ISSN: 0976 – 5999.

[4]. Ms. Swaminarayan Priya R. et al, "Knowledge Representation of Published Articles in Semantic Web using Upper Ontology", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 8, August 2012.

[5]. R. Baracho, G. Silva, and L. Ferreira, "Sentiment Analysis in Social Networks: a Study on Vehicles" ONTOBRAS-MOST, CEUR Workshop Proceedings, page 132-143, volume 938. 2012

[6]. Gopinath Ganapathy and S. Sagayaraj, "Automatic Ontology Creation by Extracting Metadata from the Source code", Global Journal of Computer Science and Technology, Vol.10, Issue 14, Ver.1.0, November 2010.